An Intelligent Learning System for Supporting Interactive Learning through Student Engagement Study

Yuanyuan Liu
Faculty of Information Engineering
China University of Geosciences
Wuhan, China 430074

Zhong Xie
Faculty of Information Engineering
China University of Geosciences
Wuhan, China 430074

Jingying Chen*
National Engineering Research Center for E-Learning
Central China Normal University
Wuhan, China 430049

***Abstract*—Interactive learning in class or off-class is crucial to teaching and learning. In this paper, we propose an intelligent learning system for supporting student interactive learning through engagement study, which is based on three modules, i.e., attendance management, teacher-student (T&S) communication, visual focus of attention(VFOA) recognition. Attendance management matches the student's identity and locates his/her profile. T&S communication provides an additional channel of Question and Answer (Q&A) between a teacher and students. VFOA recognition captures student's attention on different learning targets based on the estimated head poses, visual environment cues and prior state in class. Student engagement is analyzed based on multiple cues of one's attendance, class communication and VFOA. The experimental results suggest that an intelligent learning system is benefit to improve interactive learning and enhance learning efficiency.**

Keywords- Technology-enhanced learning; Intelligent Learning system; Engagement study; VFOA recognition

I. INTRODUCTION

Interactive learning in class or off-class is known to benefit learning and is a very important component of teaching and learning [1], [2]. In a traditional classroom, many obstacles inhibit interactions between students and teachers, such as limited class time, rigid seating arrangement and students' reservations to speak out in class [2]. Recently, technologies embedded in the learning environment can provide opportunities for educators and learners to obtain efficient learning result, e.g., WebAnn, Epost, intelligent learning system [3], [4]. However, complicated operations with PC or using invasive sensors makes these technologies hard to popularize in classroom for interactive learning. In a real and wide range classroom, intelligent and non-invasive interactive learning is still a challenging task due to complex environment and various learning progress with multi-students.

In this paper, we propose an intelligent learning system for supporting student interactive learning through engagement study. And we investigate student engagement in class based on multiple cues of his/her attendance, communication and attention to support interactive learning. Class attendance is essential to student engagement study, in this work, student attendance is recorded using face recognition. Nowadays, most students have smart cell phones and they bring them to the classroom. Students and teachers can benefit from the additional channel of communication via their cellphones in the classroom [4]. We develop a Question and Answer (Q&A) application run on cell phones to support student-teacher communication. VFOA of students in class is an important non-verbal communication cue for interactive learning and assessment for learning efficiency [5]. However, VFOA recognition remains challenging in natural environment, e.g., various illuminations, occlusions, expressions and wide scenes. In this paper, we propose a VFOA model in the natural classroom based on students' prior state, environmental cues and head poses.

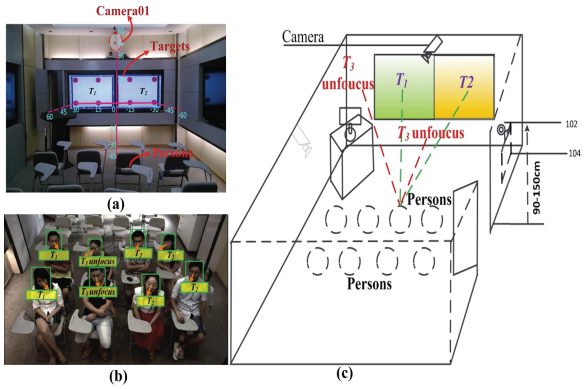


Fig. 1. The intelligent interactive learning system. (a) The real interactive learning environment, (b) The recognized results, (c) The geometric configuration of intelligent learning system.

Fig.1 shows the intelligent learning system where the student engagement is studied. The real interactive environment is given in Fig.1(a). The real class engagement states from student's attention are recognized in Fig.1(b). The geometric configuration of the intelligent learning system is shown in Fig.1(c), where Ti {T1, T2, T3} represent the VFOA of a student on left slide, right slide and outside position of the white board in the system, respectively.

The architecture of the engagement study in the intelligent learning system is given in Fig.2, which includes three modules, i.e., attendance management, T&S communication,

978-1-5090-4093-3/16/$31.00 ©2016 IEEE

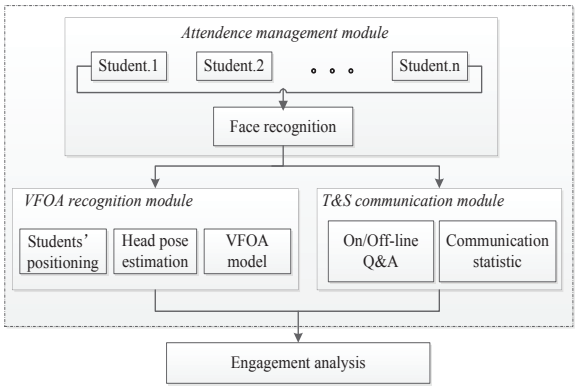


Fig. 2. Architecture of the engagement study in the intelligent learning system.

and VFOA recognition. When a student enters the class environment, attendance management matches his/her identity and locates his/her profile using face recognition with Camera at the entrance of the classroom(see Fig.1(c)). During the class, T&S communicate module assists the student and teacher interaction via the designed Q&A application run on their own cell phones and keeps all the communication records. Meanwhile, student's head poses and position are estimated to predict his/her attention using the proposed hybrid multi-layered random forest method with Camera in the system(see Fig.1(c)). Finally, multiple cues of information from three modules is fused to analyze student engagement, which assists teacher understand student's learning status and provide appropriate teaching to make learning efficient.

II. ATTENDENCE MANAGEMENT

Student and teacher should first register their information with the attendance management module, i.e., ID, name, subject, and photo. When a student enters the classroom, attendance management matches his/her identity using face recognition and extracts his/her clothing features for VFOA recognition. After identity matching, this module automatically sends each attended student's information into the teacher, and generates a unique QR code associating with each attended student to authorize him/her to access the T&S communication module. This module helps to track each attended student's learning interaction in class.

III. T&S COMMUNICATION

Different from the interaction between students and the teacher in the traditional class mode, T&S communicate module assists the student and teacher interaction via the designed Q&A application run on their own cell phones and keeps all the communication records. The student logins in the T&S communication module using the QR code provided from the attendance management module. Fig.3 shows the framework of the T&S communication. The specific steps are in the following.

During the class, a teacher can release a class topic or exercises through the on-line question function. Meanwhile,

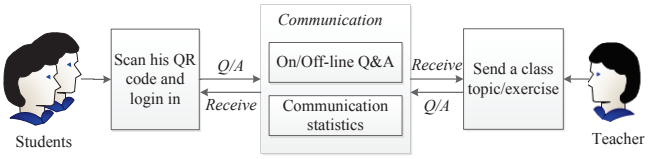


Fig. 3. T&S communication in class.

students can receive the topic or exercises and submit their answers through the application on their own cellphone with on-line answer function. Then, with the communication statistic function, students and the teacher can get the statistical analysis data of interactions, e.g., each student's answer, and the distribution of all the students' answers. This statistical data helps the teacher understand the students' learning performance and improve teaching efficiency.

In addition, students can also get the registered information of other students and the teacher in the same class. It's convenient for the students and teacher to set up a discussion group after a class.

IV. VFOA RECOGNITION

VFOA recognition module captures student's attention on different learning targets during the class (i.e., T_1, T_2, T_3, see Fig.1), which includes three stages, i.e., student positioning, head pose estimation and VFOA recognition. The flowchart of the approach is shown in Fig.4. First, student positioning associates each students identity with his/her 2D position in the video sequence via face detection, face tracking and identity matching. Then, the hybrid multilayered random forest algorithm is proposed to estimate the head pose in a hierarchical way integrating classification and regression forest, which includes five layers, i.e., D-L1 and D-L2 in horizontal direction, D-L3 and D-L4 in vertical direction, D-L5 in horizontal and vertical directions. Finally, the VFOA is recognized and tracked based on head pose, prior state and environmental cues. Environmental cues include the physical placement of targets and the participant's 3D position in the classroom. Prior state comprises the prior recognized attention state and some prior 3D cues in the classroom.

A. Student positioning

Student positioning associates each student's identity with his/her 2D position in the video sequence captured using the overhead camera (see Fig.1(c)) in the class environment. Due to the wide angles and low resolution of the image, the student positioning is difficult using face recognition alone. In this case, a hybrid approach is proposed to obtain each student's position based on his/her clothing and face images matching.

First, a cascade of boosted classifiers with Haar-like feature [6] has been trained to detect faces with our database collected from an overhead camera under various poses, illumination and occlusion. After face detection, a mean shift method similar to [7] with skin color and motion information is used to track face areas in the video sequence. Then, the Grabcut algorithm [8] is used to segment clothing areas of the students

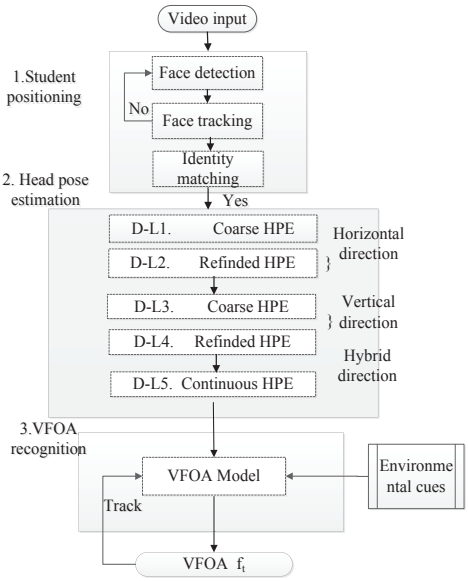


Fig. 4. The flowchart of VFOA recognition.

based on their face context information and geometric positions. Afterwards, the HOG (Histogram of Oriented Gradient) and HSV (Hue,Saturation,Value) features of the clothing area are extracted to match the stored clothing features in the attendance management using Bhattacharyya distance.

It's known that multiple students with similar clothing are difficult to distinguish the identity based on the clothing information alone, hence, we use face template matching method to further confirm the identities of students with the top 20% clothing similarity in descending order. We tested ten different video sequences with 18 students in the classroom and get the average recognition rate of 95.6%.

B. Head pose estimation

Random Forest (RF) is a popular ensemble method in computer vision because of its powerful regression and classification capability [9], [10]. We proposed a Dirichilet-tree distribution enhanced random forests algorithm (D-RF) to estimate head pose and facial features in [11], [12], [13]. In this paper, a more discriminative hybrid multilayered random forest (HMRF) based on our previous work is proposed to estimate head pose in a hierarchical way integrating classification and regression random forests. Different from the previous work, in this paper, the improved HMRF is a weighted combination of classification and regression D-RF, which estimates continues head pose based on combined texture and geometric features. More detail on our previous work can be referred to [11], [12], [13].

In order to train sub-forests in the HMRF, the training images have been divided into 4 hierarchical training subsets. First, face areas are located as described in the Student positioning section and normalized to 125*125 pixels. Then we randomly extract 200 facial patches $\{P = \{F_i, H_i^m, D_i\}\}$ from each face area in sub-sets. The patch appearance F_i is defined as multiple texture feature channels $F = \{F_i^1, F_i^2, F_i^3\}$. F_i^1 contains the gray values of the raw facial patch with dimension as 31*31. F_i^2 represents the Gabor feature based PCA of facial patches with dimensions as 35*12. F_i^3 is the histogram distributions of the patches. The channel $H_i^m = \{h_i^1, (h_i^2|h_i^1), (h_i^3|h_i^2, h_i^1), (h_i^4|h_i^3, h_i^2, h_i^1), \theta_{yaw,pitch}\}$ contains the annotated discrete and continuous angles of training subsets in different sub-layers of the Dirichlet-tree, where h_i^1 are 3 yaw angles in the first sub-layer D-L1 of the Dirichlet-tree distribution , $h_i^2|h_i^1$ are refined yaw angles refined from h_i^1 in the second sub-layer D-L2, $h_i^3|h_i^2$, h_i^1 are 3 pitch angles under condition of each yaw angle h_i^2 in the fourth layer D-L3, $h_i^4|h_i^3$,h_i^2,h_i^1 are refined angles based on the above annotated angles at leaves of the Dirichlet-tree in the fourth sub-layer D-L4. $\theta_{yaw,pitch}$ are continuous head pose angles the sub-layer D-L5. D_i is the offset vector from a patch centroid to the tip of the nose. The training procedure of each sub-forest in different sub-layers is similar to [11]. The head pose and tip of the nose position probabilities of patches $p(H_i{}^m, D_i|l) = N(H_i{}^m, D_i; \overline{H_i{}^m, D_i}, \Sigma_{H_i{}^m, D_i})$ have been stored in leaves l of the trained Hybrid D-RF trees as the Gaussian probabilistic distribution, where $\overline{H_i{}^m, D_i}$ and $\Sigma_{H_i{}^m, D_i}$ are the mean and covariance matrix of head pose and tip of the nose probabilities in the i-th layer of HMRF.

In order to obtain contiuous head pose angles, a weighted composited measure is proposed to estimate continuous head pose based on multiple probabilities in D-L5. in this paper, which is defined as,

$$\arg\max_{H}(w_m p(H^m) + (1.0 - \exp(-\frac{D_i}{\gamma}))p(D_i)) \quad (1)$$

where γ is used to control the steepness of this function, the weight $w_m = P_S/P$ that is defined as the ratio of samples' number P_S of a subset to full samples' number P in each single tree of the HMRF.

Finally, the position of the nose tip D_i and head pose angles can be detected using regression voting as Eq.(1) in the D-L5 under the condition of the estimated coarse head poses in D-L1 to D-L4.

C. VFOA recognition

The objective of this paper is to recognize the attention targets in the natural classroom from a monocular camera (see Fig.1). A novel VFOA model is proposed to recognize and track attention based on head poses, prior state and visual environmental cues as shown in Fig.5. Where $h_t^k = \{\theta_{yaw,pitch}\}$ represent the estimated head poses of person k in horizontal and vertical directions at time t, c_t represents the environmental cues currently, f_t^k and f_{t-1}^k denote the VFOA states of person k at time t and prior time t-1 .

1) Environmental cues: VFOA recognition from a monocular camera is difficult due to unknown 3D cues. In order to solve this problem, we introduce an approximated method to obtain the environmental cues $c_t(T_i, B)$ based on some prior state. $T_{i=1,2}\{T_1, T_2\}$ are the physical placement of attention targets in the white board and could be measured previously (see Fig.1). $B(x, y, z)$ is the 3D position of a person estimated

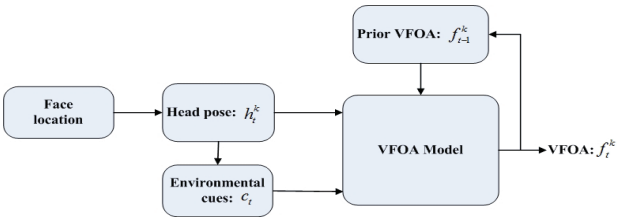


Fig. 5. The proposed model for VFOA recognition

from a monocular camera in the classroom. According to averaging 100 persons' sitting height (the height of tip of the nose) in the classroom, we fixed a 2D reference plane with its height B_y=120cm as the prior state of 3D cues. Then, 32 different points in this 2D reference plane were labelled according to their image coordinates. Homography Matrix H between the 2D reference plane and the image was computed based on these labelled points by Affine Transformation $h(\bullet)$. When recognizing, the reverse procedure could be performed, each person's position $B(x, y, z)$ can be obtained based on prior tip of the nose D_N and Homography Matrix H by Affine Transformation $h(\bullet)$.

$$B(x, y = 120, z) = h(D_N, H) \tag{2}$$

2) VFOA recognition : In the context, in order to recognize persons' VFOA in the natural classroom, the attention point $T(x, y)$ of a person is computed under the estimated head pose $\theta_{yaw,pitch}$ and his position $B(x, y, z)$. The geometric relationship between $T(x, y)$, $\theta_{yaw,pitch}$ and $B(x, y, z)$ is shown in Fig.6,

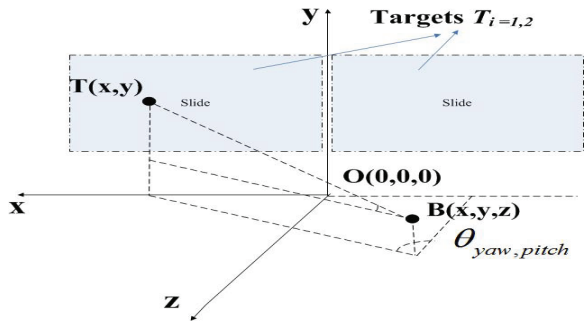
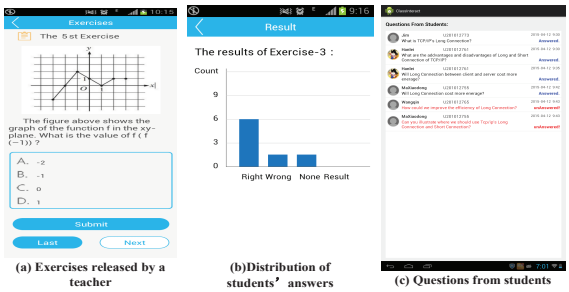


Fig. 6. Geometric relationship between the attention point, position and head pose of the student.

$$T(x, y) = \{\frac{B_z}{\cot(\theta_{yaw})} + B_x, B_y + \frac{\tan(\theta_{pitch}) \times B_z}{\cos(\theta_{yaw})}\} \tag{3}$$

If the attention point $T(x, y)$ belongs to the white board T_i , $i = 1, 2$, it means that VFOA of a person is within the white broad. Then VFOA f_t^k of a person k could be recognized as,

$$f_t^k = \sum_t \sum_{k=1, i \neq k}^{K} \delta(T_t^k - T_i), k = 1, ...K. i = 1, 2 \tag{4}$$



(a) Exercises released by a teacher (b)Distribution of students' answers (c) Questions from students

Fig. 7. Examples of T&S communication in class.

$\delta(T_t^k - T_i) = 1$ represents that a person focuses the targets T_1 or T_2. While $\delta(T_t^k - T_i) = 0$ represents that a person does not focus the double-slides, VFOA directly outputs T_3='un-focused'.

To estimate the VFOA of multi-persons of a video sequence from a classroom, we rely on GMM to track the jointed model displayed in Fig.5, and according to the jointed distribution of the estimated variables and prior state of VFOA as the VFOA model in the previous section is given by,

$$p(f_t^k | f_{t-1}^k, h_t^k, c_t) \propto p(f_0^k) \prod_{t=1}^{T} p(h_t^k | c_t) p(f_t^k | f_{t-1}^k) \tag{5}$$

where c_t is the visual environmental cue. In a video sequence, the VFOA recognition is performed by estimating the optimal sequence of states which maximizes $p(f_t^k | f_{t-1}^k, h_t^k, c_t)$.

V. ENGAGEMENT STUDY

The engagement study is performed based on multi-cues from the three modules. Decisions from each module are fused to help teacher understand the students' learning states and adjust teaching scheme. For example, according to the statistic distribution of answers from students, the teacher could get known whether the students understand the subject, otherwise the teacher may need to clarify it further. Moreover, if most of students' VFOA states are un-focused on learning targets in the white board, it could mean that the teaching slide is not attractive. The teacher may need to promote his slide contents and adjust his teaching process. The engagement study based on multiple cues of information can get a better grasp on the students' learning status which helps to improve the learning efficiency.

VI. EXPERIMENTS AND SUMMUNY

Fig.7 shows examples of the T&S communication in a class. The left image is the Q&A application on a student's cellphone, the middle image is the statistic distribution of answers submitted by students, and the right image shows questions from students received on the teacher' cellphone.

In the student positioning step, we have achieved the average detection rate of 95.6% in some public face datasets, our collected classroom database and videos. The average tracking time is 0.007s per frame on a PC with Intel(R) Core(TM) i5-2400 CPU@ 3.10GHz. The proposed approach is able to

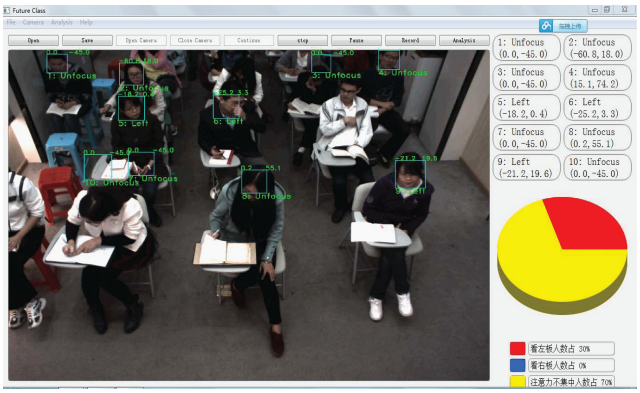


Fig. 8. An example of VFOA recognition in the system, where Left, Right, Unfocused represent the VFOA of student on the left slide, right slide and out of the white broad, respectively.

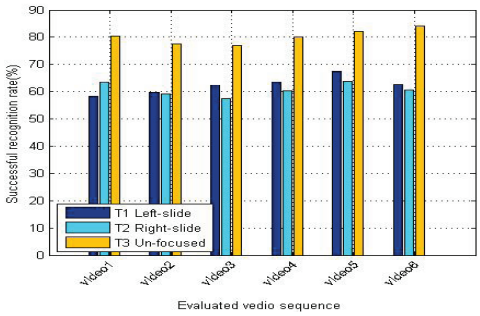


Fig. 9. VFOA recognition rate (%) in different targets.

detect tracking failures using constraints derived from the distance between persons' positions. Once the tracking failure has been detected from the camera, re-initialization and face detection is needed.

In order to evaluate multi-students' VFOA recognition, we firstly evaluated the accuracy of students' head poses in the class environment, the parameters of training and testing in HMRF are similar to [13], [14]. The average accuracy using the HMRF algorithm is 72.6% on Pointing'04 head pose dataset [15] and our collected head pose dataset from an overhead camera in the wide range classroom. Fig.8 shows an example of VFOA recognition in our intelligent learning system in real class , where the estimated head poses are shown above the face rectangles and VFOA results are below the rectangles. And the pie chart in the lower right corner of Fig.8 gives the VFOA distributions of students in class. It can help teacher understand student's learning status and provide appropriate teachin to make learning efficient.

Some recognition results are provided in Fig.9. The experiments have been performed on 6 videos in real classes. Each recording includes 8 persons in ten minutes long. The average accuracy reaches 67.8% using the proposed approach. The individual accuracy on T1, T2, T3 are 62.28%, 60.81%, and 80.2% correspondingly. The accuracies on T1 and T2 are inferior to T3 due to their smaller scales in the classroom.

Table 1 provides the VFOA recognition results obtained using different head pose estimation algorithms, including the MHRF proposed in this paper, D-RF in [11], C-RF proposed in [14], RF in [9]. One can see that the best recognition rate can be obtained using our proposed MHRF algorithms.

TABLE I
VFOA RECOGNITION RATES WITH DIFFERENT HEAD POSE ESTIMATION ALGORITHMS.

Algorithms	Recognition Rate (%)	Mean error (%)
HMRF	67.8	32.5
D-RF	63.5	38.4
C-RF	58.6	40.3
RF	50.2	46.5

The preliminary results on engagement study are based on the T&S communication process and VFOA states. Each student's questions/answers records and attention can be tracked. Teacher could improve his teaching method based on the analysis of student engagement.

VII. CONCLUSION AND FUTURE WORK

In this paper, we propose an intelligent learning system for supporting interactive learning through student engagement study, which assists teacher understand student's learning status and provide appropriate teaching to make learning efficient.

Our contributions are as follows. First, we design an intelligent learning system for investigating student engagement based on multiple cues of his/her attendance, communication and attention to improve learning interactive in class; Second, we propose a hybrid multilayered random forests for continuous head pose estiamtion in a coarse-to-fine way; Third, a novel VFOA model is proposed to recognize and track attention based on head poses, prior state and visual environmental cues.

A preliminary evaluation has been carried out in a local intelligent learning environment. The promising results suggest that the proposed method could enhance interactive learning in class and improve learning efficiency. In future, we will work towards a large-scale class study where our proposed engagement study will be carried out with more data. The impact of the interactive learning will be assessed through a range of measures including pre- and post-tests of various class sorts, along with analysis of the recorded engagement degree.

ACKNOWLEDGMENT

This work was supported by the China Postdoctoral Science Foundation (NO.2016M592406), Research Funds of CUG from the Colleges Basic Research and Operation of MOE (NO.G1323521685), National Key Technology Research and Development Program (NO.2014BAH22F01), Research Funds of CCNU from the Colleges Basic Research and Operation of MOE (CCNU16A02020).

REFERENCES

[1] Boyle J T and Nicol D J. Using classroom communication systems to support interaction and discussion in large class settings[j]. *Research in Learning Technology*, 2003.

[2] Siau K, Sheng H, and Nah F F. Use of a classroom response system to enhance classroom interactivity [j]. *Education IEEE Transactions on*, 49(3):398–403, 2006.

[3] Chen J, Luo N, Liu Y, and et al. A hybrid intelligence-aided approach to affect-sensitive e-learning[j]. *Computing*, 2014.

[4] E Scornavacca, S Huff, and S Marshall. Developing a sms-based classroom interaction system. *Proceedings of Mobile Learning Technologies and Applications (MoLTA 2007)*, pages 47–54, 2007.

[5] S O Ba and J M Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues[j]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):101–116, 2011.

[6] Viola M, Jones M J, and Viola P. Fast multi-view face detection[j]. *Proc of Computer Vision Pattern Recognition*, 2003.

[7] Yang C, Duraiswami R, and Davis L. Efficient mean-shift tracking via a new similarity measure[c]. *In IEEE CVPR*, 1:176–183, 2005.

[8] Rother C, Kolmogorov V, and Blake A. Grabcut–interactive foreground extraction using iterated graph cuts[j]. *Acm Trans Graph*, pages 309–314, 2004.

[9] L Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[10] Wang S et al. Luo C, Wang Z. Locating facial landmarks using probabilistic random forest[j]. *Signal Processing Letters, IEEE*, pages 2324–2328, 2015.

[11] Liu Y and Chen J.and et.al. Robust head pose estimation using dirichlet-tree dis-tribution enhanced random forests[j]. *Neurocomputing*, pages 42–53, 2015.

[12] Liu Y., Chen J., and Shan C. Dirichlet-tree distribution enhanced random forests for facial feature detection. *In IEEE ICIP*, pages 235–238, 2014.

[13] Liu Y., Chen J., and Chen H. Dirichlet-tree distribution cascaded hough forests for continuous head pose. *In IEEE CISP*, pages 554–559, 2014.

[14] Dantone M. and Gall J.and Fanelli G.and VanGool L. Real time facial feature detection using conditional regression forests. *IEEE CVPR*, 2012.

[15] N. Gourier and D. Hall. Estimating face orientation from robust detection of salient facial features in pointing 2004. *ICPR international Workshop on Visual Observation of Deictic Gestures*, 2004.